Scientometrics (2022) 127:6827–6853
https://doi.org/10.1007/s11192-022-04399-2

A visual analytics approach for the assessment of information quality of performance models—a software review

Marco Angelini[1] · **Cinzia Daraio**[1] · **Luca Urban**[1]

Received: 31 October 2021 / Accepted: 27 April 2022 / Published online: 4 July 2022

Abstract
In this paper we provide a review of the main functionalities of a Visual Analytics Environment (VAE) developed for the assessment of data and information quality in the context of performance evaluation of research organizations. Performing data and information quality tests are necessary procedures to ensure the bibliometric and research performance evaluation analysis of organizations have the necessary robustness. The proposed environment is helpful to guide the user to an Information Quality-aware development of Performance models. This interactive visual analytics environment offers to the user the possibility to produce and compare information quality-aware indicators, exploring and defining correct behavior, identifying anomalous cases from both data quality and information quality perspectives, and supporting the user in forming hypotheses on possible causes for those anomalies. The proposed approach, exploiting visual interactive exploration results in a more efficient process, minimizing the number of cases for which a manual investigation is needed. The illustration on European higher education institutions data demonstrates the use of the presented functionalities and their benefits.

Keywords Visual analytics · Data quality · Information quality · Research and innovation data · Higher education data

Introduction and background

In the last decade, the rapid increase in the production, communication, and evaluation of research have been signs of a transformation. This transformation has changed also the way we conceive and measure performance indicators. Generally speaking, performance is the result

✉ Marco Angelini
angelini@diag.uniroma1.it

Cinzia Daraio
daraio@diag.uniroma1.it

Luca Urban
urban.1651869@studenti.uniroma1.it

[1] DIAG Department, Sapienza University of Rome, Via Ariosto, 25 00185 Rome, Italy

(outcome) obtained by an activity and indicators are results of mathematical operations with data. When indicators are used in an evaluation process, they are called metrics. The definition and measurement of performance indicators require the formulation of a performance model., i.e., the description of the conceptual, methodological, and data dimensions capable to assess the outcome of given activities of the units under analysis, with respect to some predefined goals. An example of a performance model can be an efficient frontier model that describes how well universities produce their scientific and teaching outputs with respect to inputs and other factors (i.e., including staff, public and private funding, and so on).

Despite the various innovations introduced with big data, machine learning, and altmetrics, the role of the *user* of metrics, and her interactions in the development and evaluation phase of performance models have not received the great attention they deserve. In addition, important aspects for the usability of data and information, such as the different dimensions of *data and information quality*, have frequently been overlooked, making the developed performance measurement systems rigid, fragile and inconsistent.

Koltay (2016), discussing data quality and research data management, shows that data governance and data literacy are two important constituents to keep into account when developing a knowledge base. "Applying data governance to research data management processes and data literacy education helps in delineating decision domains and defining accountability for decision making. Adopting data governance is advantageous, because it is a service based on standardized, repeatable processes and is designed to enable the transparency of data-related processes and cost reduction. It is also useful, because it refers to rules, policies, standards; decision rights; accountabilities and methods of enforcement. Therefore, although it received more attention in corporate settings and some of the skills related to it are already possessed by librarians, knowledge on data governance is foundational for research data services, especially as it appears on all levels of research data services, and is applicable to big data (Koltay, 2016, p. 303)." Soylu et al., (2017, 2018) describe the added value of visual methods combined with ontology-based data management for query formulation and for making querying independent of users' technical skills and the knowledge of the underlying textual query language and the structure of data. Daraio et al. (2016) show the benefits of an ontology-based data integration approach for data quality in an open environment. Angelini et al. (2020) present the advantages of Visual Analytics for the development of performance models. In this paper we take a step further and present a visual analytic environment featured to performance model's development which includes data quality procedures and tests.

The paper is organized as follows. The next section describes the existing literature and how our Visual Analytics Environment bridges existing gaps to enable user interaction in the delicate stages of data and information quality. "Aim and Contribution" section illustrates the main aim and contribution of the paper. "Visual analytics environment for information quality (VAE): key functionalities" section presents the Visual Analytics Environment and its main components. "An illustration on European higher educational data" section reports an illustration on the European Tertiary Education Register (ETER) dataset, "Availability of the tool and reproducibility" section reports the information to download the software while "Concluding remarks" section concludes the paper.

Related work

Analyzing and steering Data Quality in analytical processes are very relevant activities in computer science and data analysis [e.g., Ahmed et al. (2018) for improvement of data quality in intelligent e-CRM applications, or Vielberth et al. (2021) for security incidents], where many automatic tools exist that support these tasks, be it to different degrees. Recent work by Ehrlinger et al. (2019) surveyed 667 software tools dedicated to data quality. Among other considerations, the authors report on the large heterogeneity of the tools, evidencing several limitations: (1) more than half of them work only with proprietary solutions; (2) most of them lack implementation of important Data Quality dimensions identified in the state-of-the-art literature, (3) most of them do not support comparability due to the way in which metrics are defined, and (4) they lack user interaction and exploration capabilities supported by visualization for data quality analysis (most of them only focus on usability of basic GUIs useful just to conduct the analysis and not to explore results or support hypothesis forming and testing). The visual environment we propose is an attempt to overcome the main limitations of existing software. Our proposal aims at mitigating a) limitations (1), being a freely available software; b) limitation (2) by providing a tailored methodology tested on ETER data but generalizable to other datasets, and c) limitations (4) by providing a visual environment with exploratory and analytical capabilities. For limitation (3) our solution behaves similarly to existing state-of-the-art proposals. Our environment offers the analyst a fully interactive way for testing hypotheses, evaluating results, steering the results using visual means. In this way, the accuracy of the results can be tested interactively. Few previous works exist that have explored the use of Visualization and Visual Analytics to conduct Data Quality analysis. Sulo et al. (2005) present a tool for the visual representation of Data Quality called DaVis. DaVis uses a tabular visual representation to show a dataset, highlights inaccuracies and invalid data, and shows differences between versions of a dataset. Kandel et al. (2012) propose Profiler, a visual analysis tool for assessing data quality issues. Profiler applies data mining methods to flag problematic data and assesses the data in context by automatically suggesting a multiple coordinated visualizations environment.

Liu et al. (2018) offer a literature review on Visual Analytics for Data Quality activities, and a framework for conducting data cleansing on four data types (multimedia, text, trajectories, and graphs), while Gschwandtner et al. (2014) propose a solution for data cleansing of time-oriented data, and similarly to our proposal a set of semi-automatic quality checks, visualizations, and directly editable data tables.

Cashman et al. (2021) present CAVA, a system that integrates data curation and data augmentation with the traditional data exploration and analysis tasks, enabling the user, through a visual analytics system, foraging for attributes or identifying interesting data combinations. However, this work is highly focused in data curation and data aggregation, with less emphasis on information and data quality.

do Amor Divino Lima et al. (2020) propose a visual-interactive idiom for diagnosing missing data mechanisms. The proposed solution consists of a set of visual encodings and two derived metrics that synthesize the missing data mechanisms and the uncertainty associated with this synthesis, allowing the analyst to have more confidence or choosing the right mechanism for evaluating missing values. Song et al. (2021) conducted an empirical study to understand the effect of visualizing missing values on participants' decision-making processes while performing a visual data exploration task. The study confirmed that showing data quality measures and allowing the user to consider them during data

exploration produces different behavior than not doing it, and it is mentioned as further motivation for our proposal.

Vielberth et al. (2021) propose a process-driven quality improvement approach for human-as-a-security-sensor information. However, differently from our approach, the authors just provide the process and a use case based on tabular data, without providing any means, visual or not, to support the user in this process.

Conversely, He et al. (2021) propose a Visual Analytic solution for Outliers filtering and management of information quality for automatic identification systems (AIS). The Visual Analytic solution was designed and implemented to support evaluation and exploration of AIS data quality. Finally, Bors et al. (2018) present Metrics-Doc, a visual analytic solution for assessing Data Quality that provides customizable, reusable quality metrics in combination with immediate visual feedback. This solution allows for defining specific quality metrics directly into the system, using OpenRefine syntax (OpenRefine, 2022), and test and identify data quality violations and distribution.

We highlight, as a differentiating point with respect to all these approaches, that our proposal is explicitly aimed at the evaluation of educational and research activities, considering their semantics that governs this domain and proposing specific indicators able to capture their behavior, not present in other tools.

Aim and contribution

The visual analytics environment proposed in this paper exploits Visual Analytics, "the science of analytical reasoning facilitated by interactive visual interfaces" (Cook & Thomas, 2005), focusing on the Data Quality analysis of the measures, indicators and scores that will be used by the analyst as a base for creating and developing a performance model.

Data and information quality analysis is important for bibliometric research. Bibliometric indicators are increasingly used as elements in performance evaluation systems of organizations. For this reason, the ability to assess the quality of data and information interactively and flexibly before using them in models that aim to assess performance is a relevant goal for the advancement of bibliometric research.

This is especially important given the heterogeneity of data sources, the different formats that can still convey similar semantics, and the importance that feature selection can have on the definition of a performance model.

Angelini et al. (2020) provide a workflow for dynamic creation and assessment of a performance model for evaluating research activities. This workflow is based on ontological modeling of the data sources, instantiated in the *Sapientia* ontology, that align semantically the contents coming from different sources (e.g., ETER, Scopus, Wos, and so on). From this step, the Visual Analytics Environment built allows the data exploration and builds on top of it several performance models (e.g., Efficiency models, input/output models) that can be compared and assessed to be validated. Fig. 1 shows the mentioned functionalities as steps 1 and 3.

The visual analytics environment proposed in this paper leverages on this workflow inserting a new intermediate step (see Fig. 1, step 2) that implements the evaluation of the quality of data ingested in the system, during the initial data exploration and/or once the analyst selected a pool of features on which construct the desired models (hence the bi-directional arrows for both model construction and data ingestion). For quality we mean both syntactic properties of the data, like the presence of null or incomplete values or the

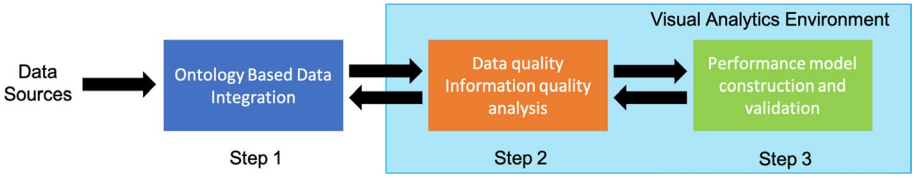


Fig. 1 Performance models development workflow

type of data at hand (e.g., categorical, numerical), and semantic properties, like the fairness of specific features (consistency) and their timeliness.

This intermediate step can reinforce the resulting quality of the developed performance models and can facilitate the check of the fairness of the model or controlling the reliability of the resulting rankings with respect to the statistical significance of the supporting data.

Given the specificity of this additional Data and Information Quality task, the resulting Visual Analytics Environment has been expanded with a tailored solution dedicated to this analysis, discussed in the next section.

Visual analytics environment for information quality (VAE): key functionalities

The proposed VAE implements a flexible approach to monitoring data quality. The core functionalities of the application are dedicated to consistency. Consistency in data quality captures the violation of semantic rules defined over (a set of) data items (Batini & Scannapieco, 2016, p. 31).

The VAE has been implemented using Streamlit, to exploit the data ingestion and management capabilities and the standard available visualization techniques. The rationale behind this choice is to focus the proposed software on user-actionable analytics without introducing custom visual solutions that could be less intuitive or require training to be understood. Additionally, this choice allows for a fast importing/exporting of analyzed data from/to the VAE (e.g., R, Stata, etc.), complementing our custom analyses with more classical ones executed externally.

Our approach is operational and empirical-based in the sense that the proposed checks do not assume any theoretical distribution for the determination of threshold parameters that identify potential outliers, inconsistencies, and errors in the data. The flexibility of our approach is that the user is at the center of the scene and has the possibility to choose the thresholds and parameters, interactively, while experimenting them. In this way, the proposed cross-sectional and multi-year controls are useful for identifying outliers, extreme observations, and for detecting ontological inconsistencies not described in the available metadata. Therefore, they can be a useful complement to the processing of available information.

The components of the VAE, are:

(a) Correlation Analysis,
(b) Map Analysis,
(c) Mono dimensional Analysis,
(d) Multidimensional Analysis,
(e) Autocorrelation Analysis,

(f) Feature Importance Analysis,
(g) Ratio Analysis,
(h) Anomalies checks,
(i) Consistency Checks
(j) Time Series Forecasting.

All the components of the VAE are illustrated in Fig. 2 (the letters given to the images are the same of the components named in the previous numbered list).

In this section we focus on the description of the two main components of VAE that are the Ratios Analysis and the Consistency Checks. "Ratio Analysis" implements a kind of *static* consistency, addressing the consistency of cross-sectional data. The "Consistency Checks" implements the *dynamic* consistency, to assess the stability of multi-annual data.

These two components represent the main added value of the VAE over the shortcomings of the other available tools. In particular, Ratio Analysis and consistency checks allow us to interactively analyze potential anomalies existing in the data, considering the heterogeneity present in the analyzed observations, and distinguishing it from the anomalies that need to be detected and eliminated.

The remaining components of the VAE, implementing more classical analyses still useful for data and information quality, are described in the Appendix.

Ratio analysis

A first and simple way to analyze the existing relationships between the variables contained in a dataset is based on ratios. The analysis of ratios made up of numerator and denominator variables that represent structural dimensions of the data considered allows us to analyze the *static consistency* existing between the dimensions (variables) of a dataset.

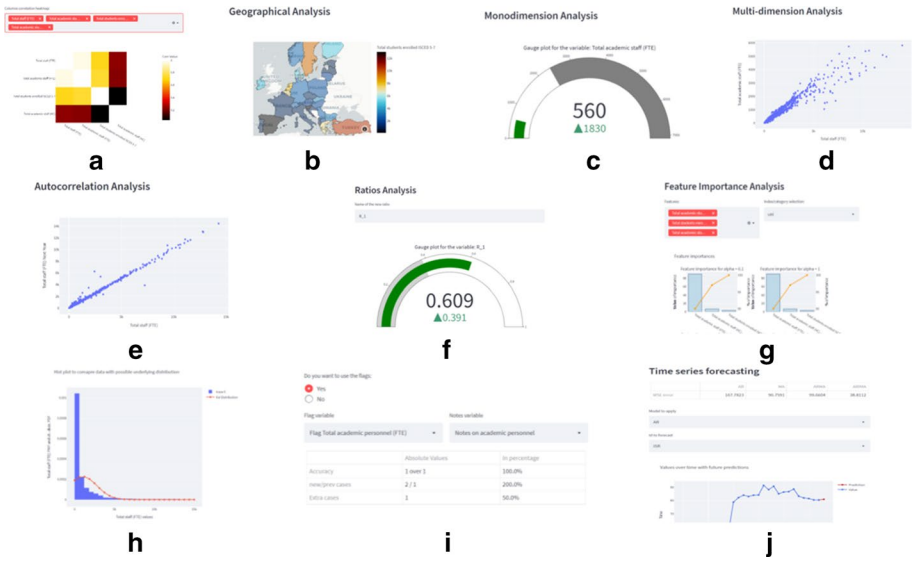


Fig. 2 Components of the Visual Analytics Environment for Information Quality (VAE)

The analysis of ratios aims at checking meaningful ratios between variables of a given dataset. In this way, it is possible to detect static inconsistencies existing between variables that have a semantic connection.

The functions contained into the Ratio Analysis are:

- *Variables ratio numerator* (multi-selection box), from which the user can choose the variables he wants to use as the numerator of the ratio.
- *Variables ratio denominator* (multi-selection box), from which the user can choose the variables he wants to use as the denominator of the ratio.
- *Name of the new ratio* (text input), from which the user can input the name she wants to give to the new created ratio; this will modify the dataset that the user can download at the end of the analysis.
- *First category col (or Nomenclature of territorial units for statistics, NUTS, id col)* (selection box) from which the user can choose a category variable or a geographical id column that will be used by the VAE to aggregate the new ratio data.
- *Second category column* (selection box) from which the user can choose a category column that will be used by the VAE to change the visualization in the violin plot. The selection of a variable in this field will produce a different violin plot for each unique category present in the chosen column.
- *Id to explore* (selection box) from which the user can choose a specific category from the ones present in the column chosen in the "**First category col (or NUTS id col)**" selection box. Choosing a specific category modifies the distribution of the created ratio only for the selected category, shown using a violin plot.

After the user selected the desired inputs, the VAE will return: a gauge plot with the distribution of the created ratio (see Fig. 3), a geographical heatmap plot (see Fig. 4) with the mean of the distribution [if the user selected a geographical id column in the "First category col (or NUTS id col)" selection box], a violin plot with the distribution for all entities and categories (see Fig. 5) and a download button from which the user can download the modified dataset (the dataset he imported plus the created ratio).

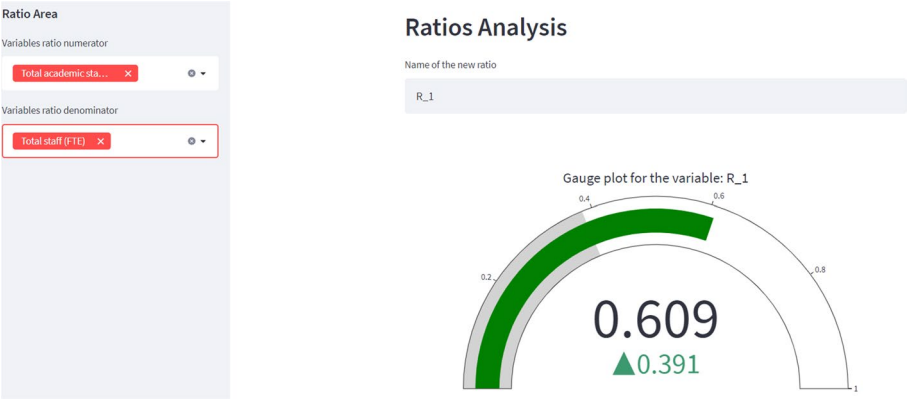


Fig. 3 Gauge Plot for Ratio Analysis

Fig. 4 Geographical heatmap Plot for Ratio Analysis

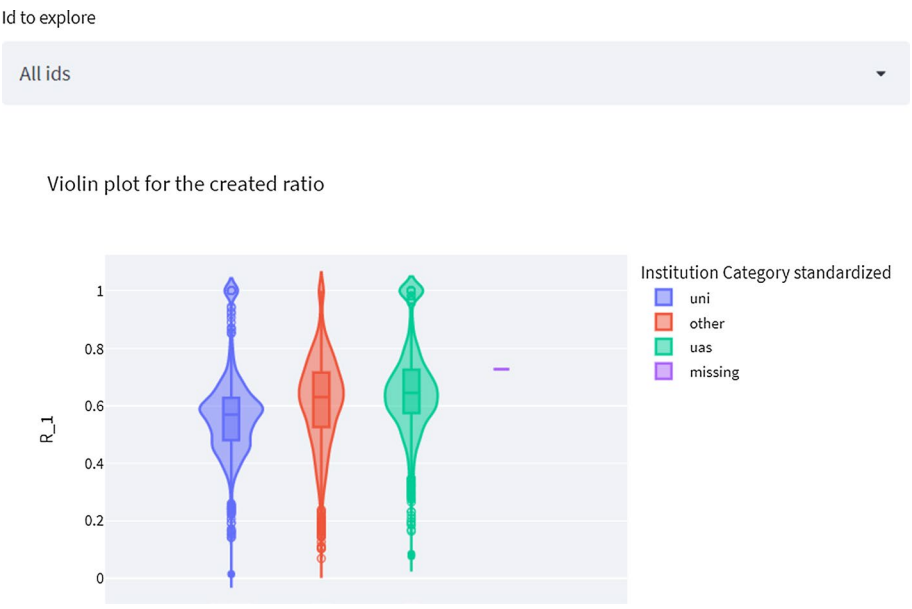


Fig. 5 Violin Plots for Ratio Analysis

Consistency checks

The objective of consistency checks is to identify anomalies present in a dataset by considering the semantics of the variables and the relationships existing between variables. An important part of the consistency analysis is represented by the dynamic consistency that aims to analyze the evolution of the consistency of variables and structural relationships between variables (e.g., ratios) over time.

Consistency checks in the VAE software is based on a multiannual stability analysis and can be run also on ratios.

The Multiannual Analysis allows to spot interesting behaviors with respect to the selected feature. It allows to explore the selected subset of features (from filters present in the left part of the VAE) and spot potential correlations, similar trends, or outlying elements to further investigate. Daraio et al. (2020) detail the application of this approach for higher education data.

The functions contained in the Consistency Checks are summarized in Fig. 6 and are the following:

- *Analysis to apply* (selection box), from which the user can choose the analysis that she wants to apply (can be "Multiannual Analysis" or "Ratio Analysis"). This choice will modify the layout of the page, the calculations that will be applied and the produced results.
- *Index col* (selection box), from which the user can choose the variable that will be
- used to select the different entities in the dataset, affecting the data aggregation.
- *Country col* (selection box), from which the user can choose the variable that contain the geographical code. This will change the aggregation and the choice (only in the Ratio Analysis) of the flagged cases, so the system at the end of the computation can give a more specific view of how the flagged cases are geographically distributed and where these "problematic" cases are concentrated.
- *Category col* (selection box), from which the user can choose a category variable that will be used from the application to compute the cases to flag (only in the Ratio Analysis) and to split the visualization results by the unique values present in the chosen column for the chosen entities, so with this selection the user can have a more complete representation of the check's results and how the flagged cases are distributed.
- *Quantile to exclude from the calculation (S1)* (numerical input) from which the user can parametrize the numerical threshold value (must be between 1 and 10) that will be used as quantile to exclude the entities that have very low values for the selected

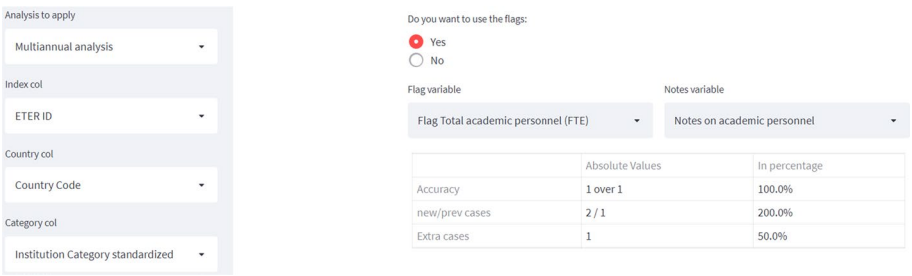


Fig. 6 Multiannual checks analysis setup

variable. This helps prevent meaningless results (this input will not be present if the user choses the *Ratio Analysis* in the "**Methodology to apply**" input selection box).

- *Flags quantile (S2 and S3)* (numerical input), from which the user can parametrize the numerical value (must be between 35 and 100) that will be used as a threshold quantile to select the entities that will be considered "*problematic*" by the application.
- *Percentage problematic cases* (numerical input), from which the user can parametrize the numerical value (must be between 0 and 100) that will be used as a percentage threshold to select the "*problematic*" countries and/or categories. These "*problematic*" cases will be treated in a separate analysis.
- *p-value percentage trend estimation* (numerical input), from which the user can define the numerical value (must be between 5 and 50) that will be applied as *p*-value to amplify the rejection region of the Mann–Kendall test Hussain et al., (2019), to distinguish between weak and undetermined trends (can be useful when the user has to analyze a time series with few values).
- *Variable consistency checks* (selection box), from which the user can choose the variable on which the checks will be applied. This will affect the results and the flags that will be issued by the system.
- *Do you want to use the flags* (radio button), from which the user can choose if she wants to use the flag column or not. Using known flags will allow a validation of the identified anomalies threshold, opposed to not using them (they act as a pre-known ground-truth).
- *Flag variable* (selection box), from which the user can choose the variable that contains the flags for the consistency checks. This column will be used to define which variable will be used for ground-truth by the system.
- *Notes variable* (selection box), from which the user can choose the variable that contains the notes related to the flag chosen variable. This will affect the "*problematic*" entities categorization, because in this case the VAE will distinguish between two types of *problems*, the "checked but not verified" problems and the "checked and verified" ones.
- *Institution trend type* (selection box), from which the user can choose the trend type she wants to inspect better; this selection works as a filter for the entities that will be shown in the next selection box, based on the chosen temporal trend.
- *Institution to visualize* (selection box), from which the user can choose the entity that she wants to focus the trend analysis. In this way the user can confirm the categorization that the application made or modify it by changing the *p-value* with the "***p-value percentage trend estimation***" function described earlier.
- *Do you want to compare trends?* (radio button), from which the user can choose if she wants to visually compare the trends of two different inspected variables for the same flagged entities.
- *Variable to compare* (selection box), from which the user can choose the variable that will be used to make a trend comparation between this variable and the one that the user previously selected in the "**Variable consistency checks:**" selection box. With this field the user can see if the same entity has an equal or an opposite trend for these two variables, this can help him in the entity analysis.
- *Variable time values* (selection box), from which the user can choose the variable that contains the different time values present in the dataset. These values will be used by the VAE to order the time values contained into the dataset and create a new column

for each one of them in the result dataset that the user can download at the end of the tool page.

- *Descriptive columns* (multi-select box), from which the user can choose the variables that she wants to add to the final dataset. The resulting dataset can be downloaded by the user at the end of the analysis.

If the user selects a flag column in the "Flag variable" selection box a table will appear in the results (like in the example presented in Fig. 6) from which the user can see the "performance" of the selected thresholds using the chosen column as a ground-truth. In this table there will be three types of metrics: first the accuracy of the new flags with respect to the already flagged ones (the number of entities correctly flagged with respect to the number of the known flagged entities used as ground-truth), then (Fig. 7) there is the number of institutions flagged with respect to the flagged entities contained in the flag variable; finally, the last metric shows the number of entities flagged that were not flagged in the chosen column (e.g., new elements to check or potentially false positives).

	CH	Total
Total academic personnel (FTE) (university)	2 (16.67%)	2 (16.67%)
Total academic personnel (FTE) (university of applied sciences)	0 (0.0%)	0 (0.0%)
Total academic personnel (FTE) (other)	0 (0.0%)	0 (0.0%)
Total	2 (5.56%)	2 (5.56%)

	Variable	Country	Category	% Value	Absolute values
empty					

	Number of institutions
Strong decrease	0
Weak decrease	0
Undetermined trend	2
Weak increase	0
Strong increase	0

	Absolute Values	In percentage
Accuracy	1 over 1	100.0%
new/prev cases	2 / 1	200.0%
Extra cases	1	50.0%

Fig. 7 Breakdown of cases by countries, categories and trend classification

Institution trend type

Undetermined trend

Institution to vizualize

CH0010

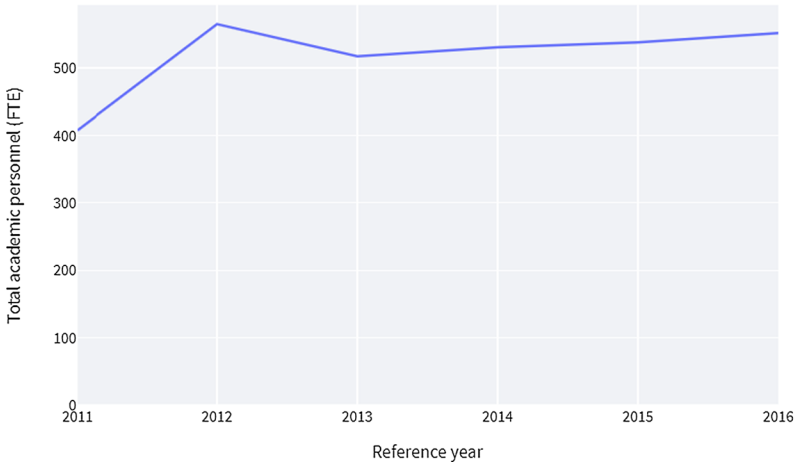


Fig. 8 Trend visualization

Variable to compare:

Total personnel (FTE)

	(Total academic personnel (FTE)) Increasing	(Total academic personnel (FTE)) Unknown	(Total academic personnel (FTE)) Decreasing
(Total personnel (FTE)) Increasing	0	0	0
(Total personnel (FTE)) Unknown	0	2	0
(Total personnel (FTE)) Decreasing	0	0	0

Fig. 9 Trend comparison table

At the end of the page there will be a button named "Download data with labels" from which the user can download the dataset for reporting activities or continuing the analysis with different tools (Figs. 8, 9).

An illustration on European higher educational data

The system has been designed and developed for broad and general use. The VAE may be particularly useful for science and innovation databases such as those included in the Research Infrastructure for Science and Innovation Policy Studies (RISIS, for more information see: https://www.risis2.eu/risis-datasets/).

To illustrate the VAE, we use the data on European Higher Education Institutions (HEIs) for which specific ratios and multiannual checks were developed (see Daraio et al., 2020), but other ratios could be introduced or created in the system without loss of generality. The European Tertiary Education Register (ETER) recalled above is a database supported by the European Commission that collects information on individual European Higher Education Institutions (HEIs). It provides data on their basic characteristics and geographical information, number of students, graduates, international doctorates, staff, fields of education, income and expenditure. ETER data are currently available on the project website (https://www.eter-project.com/) and includes 2964 HEIs from EU-27 countries and Albania, Iceland, Liechtenstein, Montenegro, Norway, Serbia, Switzerland, Turkey, UK and the Republic of North Macedonia. The available data cover the period 2011–2016.

This section briefly describes how to perform multi-annual checks on a dataset downloaded directly from the ETER database. To import the ETER dataset:

(1) open the application at: https://share.streamlit.io/lucaurban/visual_analytics_environment/main/appVAE.py
(2) select the "ETER Dataset" option
(3) Click on the "Browse files" button to upload the dataset you previously downloaded from the ETER website.

Multiannual checks

The main steps to run a multiannual check analysis are the following:

(1) First the user chooses:

   (a) the id variable ("ETER ID")
   (b) the country id ("Country Code")
   (c) the category variable (for the multiannual check can be omitted but for the ratio check it is mandatory, "Institutional Category standardized" in this example)
   (d) the name of the variable on which to issue the flags ("Total academic personnel (FTE)").

(2) After that, the user can choose if use the flags (ground-truth) if available and the notes on the flags (this field is not mandatory).

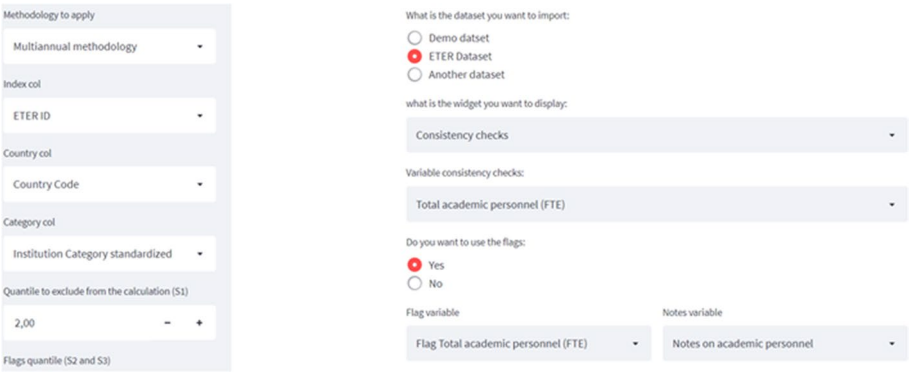
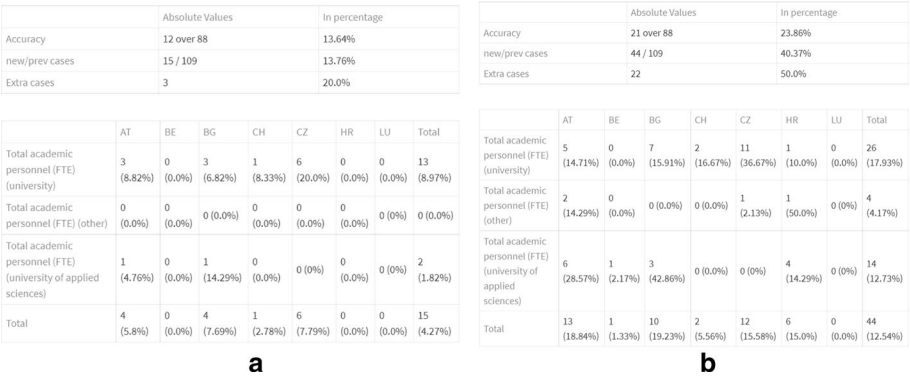


Fig. 10 Example of configuration page for Consistency Checks on ETER dataset



	Absolute Values	In percentage
Accuracy	12 over 88	13.64%
new/prev cases	15 / 109	13.76%
Extra cases	3	20.0%

	AT	BE	BG	CH	CZ	HR	LU	Total
Total academic personnel (FTE) (university)	3 (8.82%)	0 (0.0%)	3 (6.82%)	1 (8.33%)	6 (20.0%)	0 (0.0%)	0 (0.0%)	13 (8.97%)
Total academic personnel (FTE) (other)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0%)	0 (0.0%)
Total academic personnel (FTE) (university of applied sciences)	1 (4.76%)	0 (0.0%)	1 (14.29%)	0 (0.0%)	0 (0%)	0 (0.0%)	0 (0%)	2 (1.82%)
Total	4 (5.8%)	0 (0.0%)	4 (7.69%)	1 (2.78%)	6 (7.79%)	0 (0.0%)	0 (0.0%)	15 (4.27%)

a

	Absolute Values	In percentage
Accuracy	21 over 88	23.86%
new/prev cases	44 / 109	40.37%
Extra cases	22	50.0%

	AT	BE	BG	CH	CZ	HR	LU	Total
Total academic personnel (FTE) (university)	5 (14.71%)	0 (0.0%)	7 (15.91%)	2 (16.67%)	11 (36.67%)	1 (10.0%)	0 (0.0%)	26 (17.93%)
Total academic personnel (FTE) (other)	2 (14.29%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (2.13%)	1 (50.0%)	0 (0%)	4 (4.17%)
Total academic personnel (FTE) (university of applied sciences)	6 (28.57%)	1 (2.17%)	3 (42.86%)	0 (0.0%)	0 (0%)	4 (14.29%)	0 (0%)	14 (12.73%)
Total	13 (18.84%)	1 (1.33%)	10 (19.23%)	2 (5.56%)	12 (15.58%)	6 (15.0%)	0 (0.0%)	44 (12.54%)

b

Fig. 11 Example of flags obtained by the Multiannual Analysis of Consistency Checks

(3) After completing the desired configuration, the VAE will produce the results shown in the next figures (Fig. 10–13).

Figure 10 shows an example of the basic configuration you need to select to produce the results from the VAE application.

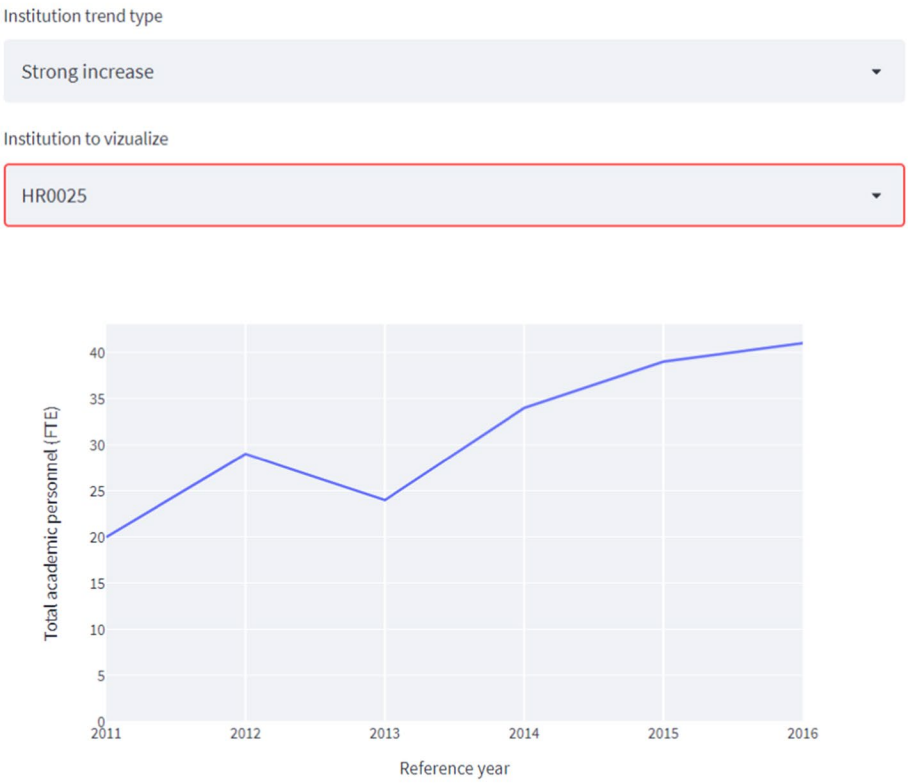
Figure 11 shows two tables with results from the Multiannual Analysis.

Figure 11a shows the result of a comparison with previously specified existing flags. This table will not be shown in the case the user chooses the option "No" related to the flags in the VAE. Table (a) and Table (b) of Fig. 11 show the results obtained with two different quantile flags [95th for Table (a) and 85th for Table (b)]. By changing the quantile threshold, the user can tune this analysis. Selecting the 95th quantile the user flags a very low number of institutions, just 15 institutions, with respect to the 109 institutions included in the ground-truth [see the second row of Table (a)]. Of these 15 institutions, 12 were flagged as problematic cases also in the ground-truth [as reported in the first row of Table (a)] while there are three institutions that were not considered previously [reported as "extra cases" in the third row of Table (a)]. If the user changes the quantile to the 85th [Table (b) of Fig. 11] she obtains a higher number of flagged institutions: 44 against the 15 previously flagged. Among these 44 institutions, 21 were flagged as problematic cases

	Number of institutions
Strong decrease	1
Weak decrease	2
Undetermined trend	36
Weak increase	2
Strong increase	2

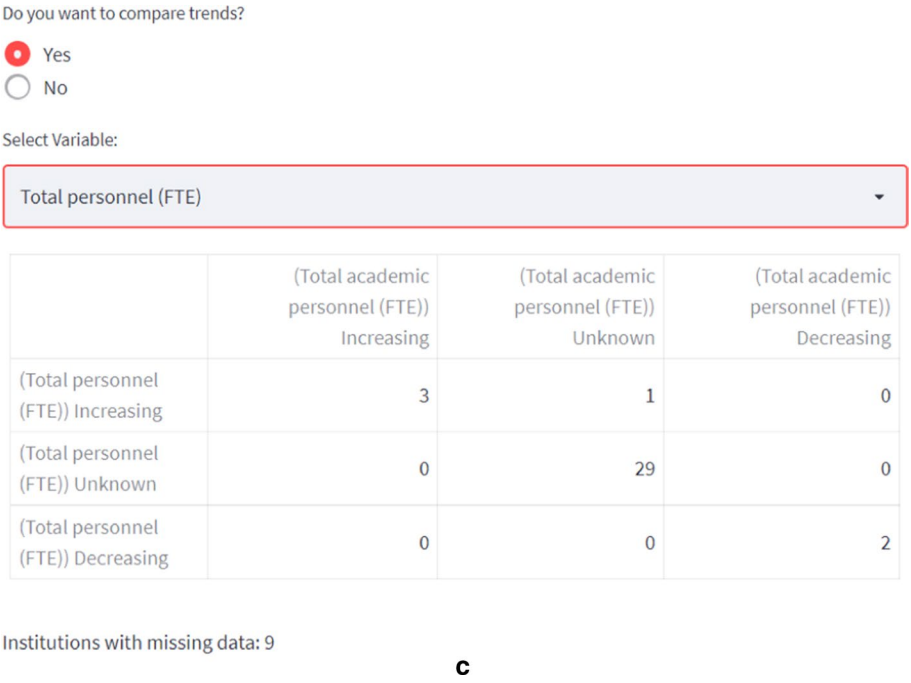
	Absolute Values	In percentage
Accuracy	20 over 88	22.73%
new/prev cases	39 / 109	35.78%
Extra cases	19	48.72%

a



b

Fig. 12 Results of the trend analysis using the 85th quantile



c

Fig. 12 (continued)

in the ground-truth, 22 of them were not considered previously and one institution was checked but there were no problems into it.

The tables at the bottom of Fig. 11 show the flagged institutions aggregated by country (AT, BE, BG, CH, CZ, HR, LU) and by category (university, other and university of applied sciences).

Figure 12 shows the results of the *trend* analysis obtained selecting the 85th quantile threshold.

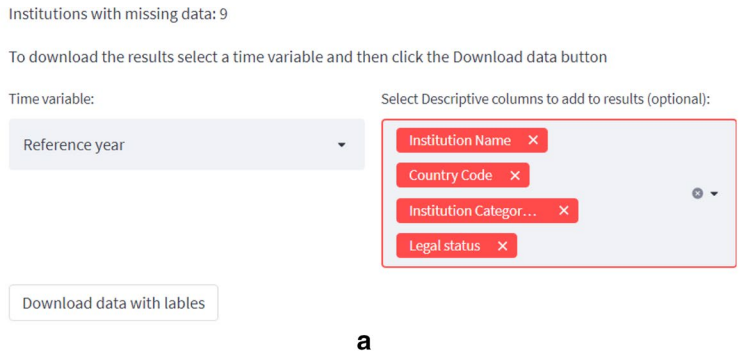
In Fig. 12a, the first table shows the classification by trend of the flagged institutions (in this example, the 2 flagged institutions are classified as "Undetermined trend"), while the second table compares the predefined flagged institutions with the institutions flagged by the application (in this analysis) that have *strong increase*, *strong decrease* or *undetermined trends*. The number of institutions that will be flagged after this analysis is 39 against the 44 previously flagged and from these 39 institutions: 20 were flagged as problematic cases in the ground-truth (1 less respect the initial flags) and 19 were not detected previously (so are potential new cases to check). If we exclude the institutions that have a weak trend (meaning that their strange behaviour could be due to error in data collection), we slightly reduce the number of flagged institutions (39 against 44) and most of them (three up to five) were institutions that were not considered previously, one was flagged in the ground-truth and one was controlled but it was not considered problematic.

After the analysis of these results, the user can visualize the data regarding a specific institution for the variable chosen (Fig. 12b) to see if the institution was classified correctly and eventually increase or decreasing the p-value for the weak trends detection.

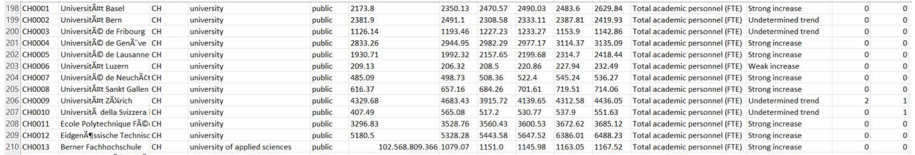
She can confront the trends of the chosen variable [e.g., Total academic personnel (FTE) in this example] with respect to another [e.g., Total personnel (FTE) in this example] for the flagged institution to see if there are concordant or discordant trends for these institutions (see Fig. 12c). In this example, from the 44 flagged institutions, for five of them the trends are concordant (three have an increasing trend for both variables while two have a descending trend), 30 of them have an undetermined trend for at least one variable and for the remaining nine it was not possible to identify a trend due to lack of data. So, the user can conclude that there are not institutions with a suspicious combination of trends.

After the completion of the analysis, the user can export and download the results in a csv file. To do this she has to choose first the time variable (e.g., Reference year) and then additional descriptive variables (not mandatory), as shown in Fig. 13a. She then downloads the resulting file, as illustrated in Fig. 13b.

Identifying and analysing outliers is a very tricky matter. Particularly in the context of university assessment, there is an enormous degree of heterogeneity across organizations. "Black box" statistical approaches, in which the user has the final result without knowing how it was obtained, can identify as outliers' data that are correct but reflect the heterogeneity present in the organizations being analysed. The approach we implemented in VAE attempts to balance the trade-off between preventing the user from relying on poor data and preventing users from employing often arbitrary data cleaning techniques (which at worst clean unusual but real data). Having the ability to choose the quantiles by which to flag in the system and to interactively analyse the results obtained, as illustrated above



a



b

Fig. 13 Download of results and export in Excel format

in Figs. 11–13 represents an empirical method of anomaly analysis and detection that is more accurate than the black box systems implemented in existing software discussed in "Related work".section. Moreover, the description of the (potential) anomalies by country and categories of institutions (university, other, university of applied sciences) allows the user to see immediately if there are problems in some specific categories of the analysed institutions.

Availability of the tool and reproducibility

The proposed software has been implemented and tested on a case study based on European Higher Education Institutions data (ETER) and is freely available (under GNU General Public License v3.0, https://www.gnu.org/licenses/gpl-3.0.html) for further testing and extension to other datasets at the following URL:

https://share.streamlit.io/lucaurban/visual_analytics_environment/main/appVAE.py.

The source code of the application can be found at the following URL: https://github.com/LucaUrban/visual_analytics_environment/blob/main/appVAE.py.

The proposed Visual Analytics Environment for evaluating Data and Information quality can be applied to any dataset or system of variables, including the datasets of the RISIS infrastructure.

The software will be kept maintained and up to date with respect to technology aspects, with the possibility that new features and analyses will be added, resulting by the current and future activities involving the VAE.

We plan to expand the reproducibility support by adding the possibility to automatically log the user's parameterizations and triggered functions, allowing to export a script capable to be replayed automatically for verification purposes.

Concluding remarks

This paper proposes a review of the main functionalities of a Visual Analytics Environment devoted to assessing the data and information quality of complex datasets, characterized by a high heterogeneity of the main dimensions. The Visual Analytic approach that the software allows is able to reinforce the quality analysis of the information that can be subsequently considered in a performance evaluation model. The Visual Analytic approach we propose facilitates the checks on the distributions and variability of data and carries out a selection of the variables and units prior to the development of performance models. It additionally facilitates the identification of the anomalies present in the data and helps to identify their potential causes. It offers an opportunity for improving data quality-aware empirical investigations.

The data-driven quality checks implemented in the VAE software are openly available and can be applied to different datasets for building and monitoring the data quality of new or existing databases, allowing the interaction of the users in defining and testing the main thresholds and parameters.

The illustration on European universities microdata, characterized by a high heterogeneity of types and categories of HEIs across countries, shows the usefulness of our

visual environment, its flexibility and the advantages of an interactive visual analytic environment over existing black box software.

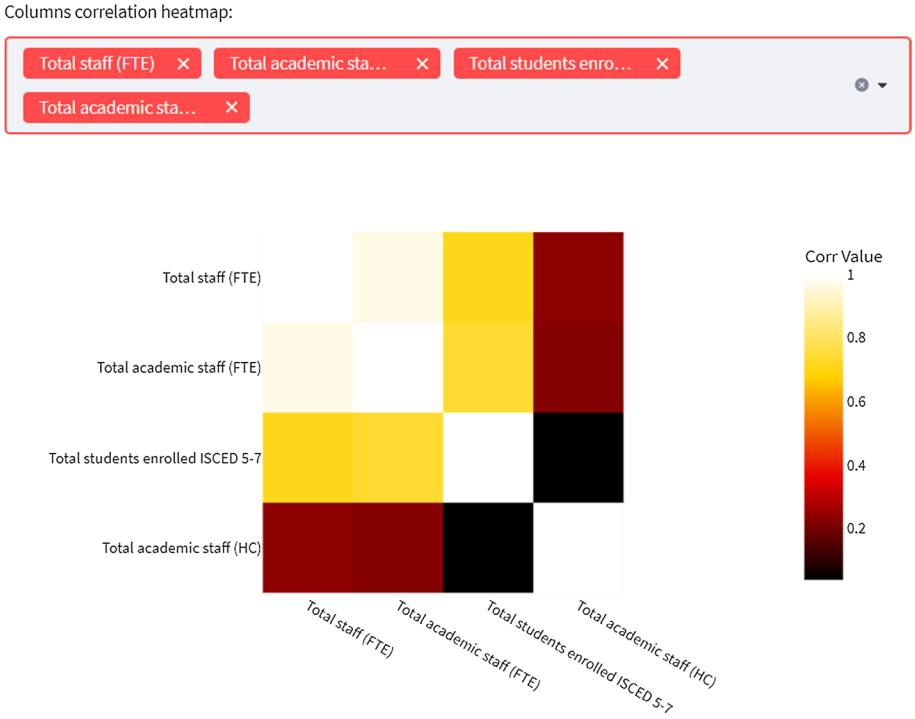
A strength of the VAE is its empirically oriented flexibility that allows the user to customize the parameters with respect to the observed distribution of the variables considered instead of using theory-based distribution functions for the data being analyzed. The VAE combines multi-dimensional controls and cross-sectional checks to further reduce the number of cases to be manually inspected and to pre-identify problems or possible explanations.

Appendix

Correlation analysis

The functions contained in the Correlation Analysis are:

- *Columns correlation heatmap* (multi-selection box) from which the user can choose the columns that will be taken by the application to calculate the correlation between them and to create the heatmap that will contain this information like in the next example.

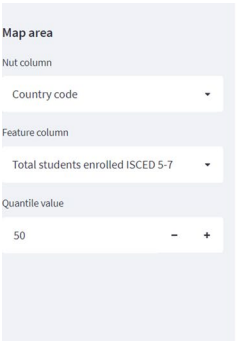


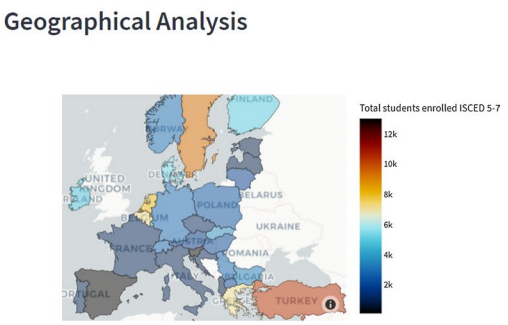
Geographical analysis

The functions contained into the Geographical Analysis are:

- *NUTS column* (selection box) from which the user can choose the variable that contains the geographical code, this will be used by the application to know the type of areas he has to plot in the result and the aggregation areas.
- *Feature column* (selection box) from which the user can choose the variable from which the system will extract the values that will be plotted in the map plot.
- *Quantile value* (numerical input) from which the user can choose the quantile value (integer value from 1 to 100) that will be calculated from the variable previously chosen, this will affect the value and the color that will be shown in the result.

Using these inputs functions the application will produce a colored map plot. The user can interact with this map by focus/de-focus or sliding the view and he can see the geographical code and the quantile value calculated by the application for the specific area by pointing a colored area of the map.



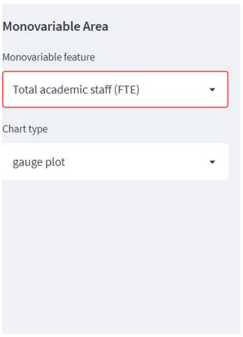


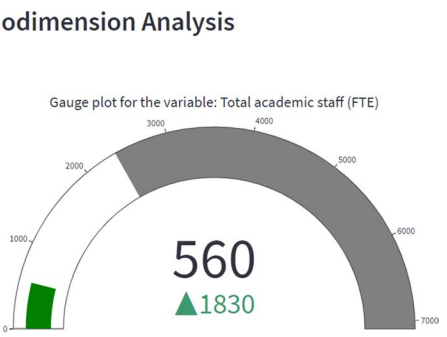
Mono dimensional analysis

The functions contained into the Mono dimension Analysis are:

- *Mono variable feature* (selection box) from which the user can choose the variable that will be used by the application to calculate the values (like the mean and the quantiles or the percentages) that will be presented in the result plot.
- *Chart type* (selection box) from which the user can choose the type of chart he wants to be produced as result (for the moment there are only the "Gauge plot" and "Pie chart" as options).

In the example presented below there is a gauge plot for the Total personnel (FTE) variable and the green bar represents the mean while the light, dark grey bars represent respectively the first and last 5° quantile of the distribution and the green number represent the difference between the mean and the 95° quantile.

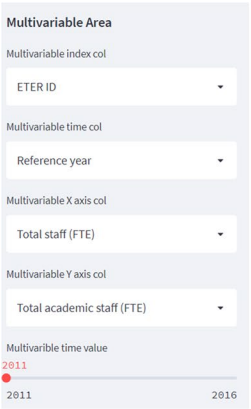


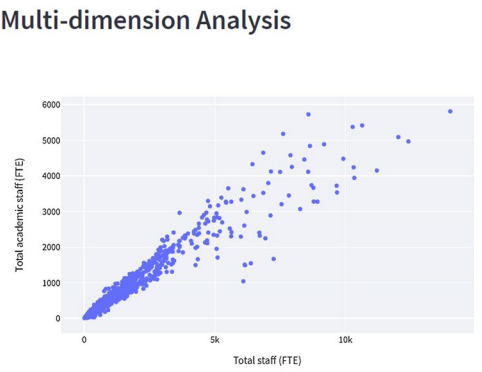


Multi-dimensional analysis

The functions contained into the Multi-dimensional Analysis are:

- *Multivariable index col* (selection box) from which the user can choose the variable that will be used to select the different entities in the dataset, this will affect the presentation of the results and the choices that the user can make in the selection box input function after "**Id time control charts**".
- *Multivariable time col* (selection box) from which the user can choose the variable that contains the different time values, if the value chosen is different from the starting value "-" (in this case all the values in the dataset will be plotted) a slider from which the user can choose a specific time value will appear.
- *Multivariable X axis col* (selection box) from which the user can choose the variable from which the tool will extract the values that will be plotted in the scatterplot on the X axis.
- *Multivariable Y axis col* (selection box) from which the user can choose the variable from which the tool will extract the values that will be plotted in the scatterplot on the Y axis.
- *Multivariable time value* (slider) from which the user can choose the time values he wants to be plotted in the scatterplot, so the user can have a different result for each different time value in the dataset.
- *Id time control charts* (selection box) from which the user can choose a specific entity from the ones contained into the variable chosen in the selection box after "**Multivariable index col**" this will affect the time control charts that will be showed by the application after this selection box.

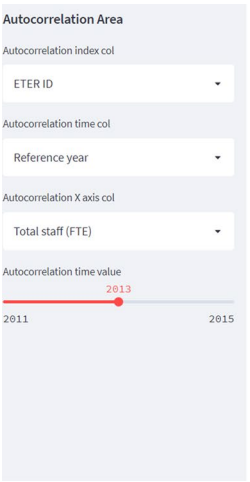


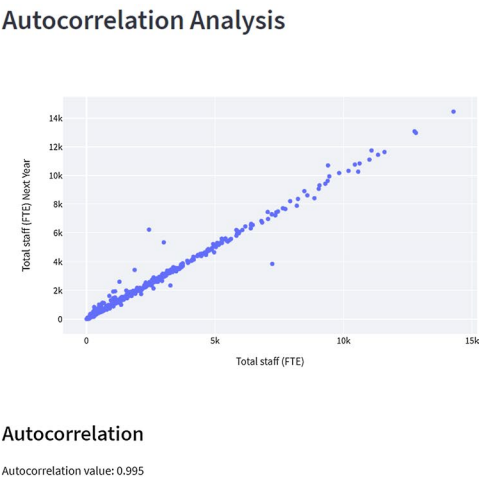


Autocorrelation analysis

The functions contained into the Autocorrelation Analysis are:

- *Autocorrelation index col* (selection box) from which the user can choose the variable that will be used to select the different entities in the dataset, this will affect the presentation of the results and the choices that the user can make in the selection box input function after "**Id deltas timeseries**".
- *Autocorrelation time col* (selection box) from which the user can choose the variable that contains the different time values, then a slider from which the user can choose a specific time value will appear.
- *Autocorrelation variable col* (selection box) from which the user can choose the variable from which the tool will extract the values that will be plotted in the scatterplot on the X and Y axes.
- *Autocorrelation time value* (slider) from which the user can choose the time values he wants to be plotted in the scatterplot; the tool will show on the scatterplot the data of the chosen time value respect the data that refers to the next time value. In the example is shown the scatterplot for the "Total staff (FTE)" variable for the 2011 respect the data of 2012.
- *Id deltas timeseries* (selection box) from which the user can choose a specific entity from the ones contained into the variable chosen in the selection box after "**Autocorrelation index col**" this will affect the time control chart that will be showed by the application after this selection box.

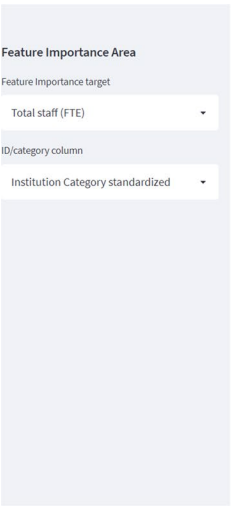




Features importance analysis

The functions contained into the Feature Importance Analysis are:

- *Feature Importance target* (selection box) from which the user can choose the variable that will be used as target by the *Ridge Regression*.
- *ID/category column* (selection box) from which the user can choose the variable that will be used to select the different entities/categories in the dataset, this will affect the choices that the user can make in the selection box input function after "**Index/category selection**".
- *Features* (multi-selection box) from which the user can choose the features that will be extracted and adapted (the data will be imputed if there're null values and scaled) to apply the *Ridge Regression* and from the result of this regression the tool will extract the *Feature Importance* of each variable to construct the *Pareto Chart* like in the example below.
- *Index/category selection* (selection box) from which the user can choose the specific entity/category on which he wants to apply the *Ridge Regression* and obtain the results.

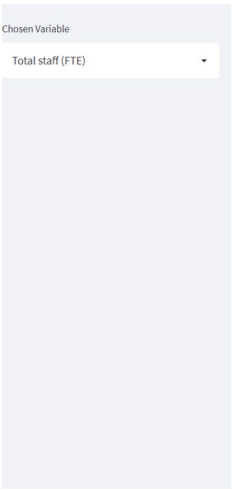


Check of anomalies

The functions contained into the Anomalies check are:

- *Chosen variable* (selection box) from which the user can choose the variable on which the application will calculate the Tukey's fences (Tukey, 1977) to find the outliers.
- *Distribution anomalies estimation* (selection box) from which the user can choose the distribution that in her opinion fits better the data from the: Normal, Exponential, Log-normal, and two-parameter Weibull distributions. Based on this choice the tool will change the formula in the calculation of Tukey's fences.
- *Tukey's constant value* (numeric input) from which the user can write the numerical value (must be between 1 and 5) that will be used to calculate the Tukey's fences with the formulas described into the page.
- *Outlier index col* (selection box) from which the user can choose the variable that contains the index related to the entities contained into the dataset, this will affect the results that will be shown at the end of the page and the way they will be aggregated.
- *Outlier type* (selection box) from which the user can choose the outlier type he wants to inspect better by seeing their concentration by country with the colored map plot that will be shown after this selection box.

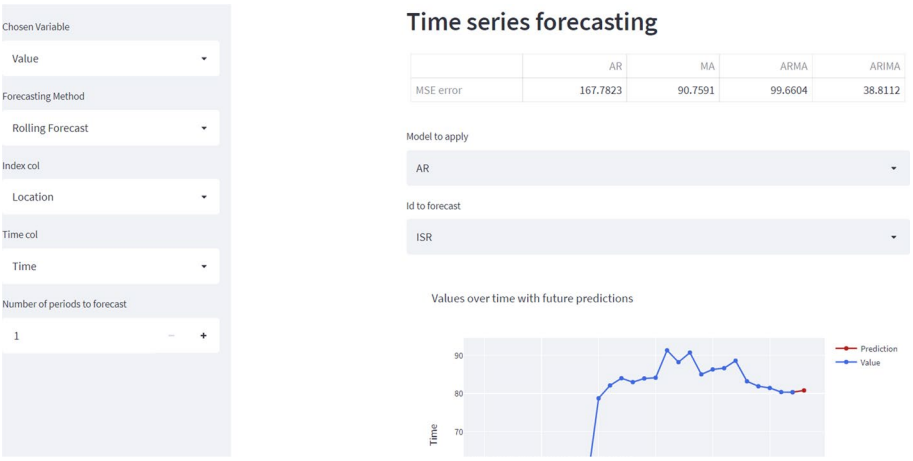
At the end of the page there will be also a table on which the user can see only the entities that have an outlier (anomaly) for the chosen variable, in this way the user can identify better the common characteristics of these entities.



Time series forecasting

The functions contained into the Time Series Forecasting part are:

- *Chosen variable* (selection box) from which the user can choose the variable on which the application will apply the time series models (AR, MA, ARMA and ARIMA), this will affect the results showed by the tool at the end of the calculations.
- *Forecasting method* (selection box) from which the user can choose the type of forecasting method he wants to apply to the entities in the dataset (the user can choose from the Rolling and Recurring method).
- *Index col* (selection box) from which the user can choose the variable that will be used to select the different entities in the dataset, this will affect the presentation of the results and the choices that the user can make in the selection box input function after "**Id to forecast**".
- *Time col* (selection box) from which the user can choose the variable that contains the different time values present in the dataset, this value will be used by the application to change the dataset structure in a simpler way to apply the models.
- *Number of periods to forecast* (numerical input) from which the user can choose the number of periods he wants to forecast (it must be an integer value from 1 to 10); this will affect the results that will be presented in the line plot at the end of the page.
- *Model to apply* (selection box) from which the user can choose the model he wants to apply (AR, MA, ARMA and ARIMA) for the estimation of the chosen variable on the entity that he will choose in the selection box after "**Id to forecast**". This choice will affect the results presented in the line plot at the end of the page.
- *Id to forecast* (selection box) from which the user can choose the entity on which he wants to apply the chosen model, the application basing on this choice will make a new data frame that will contain only the information about the chosen entity.



Acknowledgements The funding support of Sapienza Awards n. PH11715C8239C105 and n. RM11916B8853C925 is gratefully acknowledged, together with the RISIS 2, Grant agreement N° 824091, European Union H2020 Project. This paper is a substantially extended version of a research in progress paper (Angelini, Daraio and Urban, 2021) presented at the 18th International Conference on Scientometrics & Informetrics (ISSI2021), 12–15 July 2021. A preliminary version of this environment was used during a Methodological Course on Data Quality organized within the training activities of the EU RISIS Project (Research Infrastructure for Science and Innovation Policy Studies from the 15th to 17th of September 2020.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Declarations

Conflict of interest The second author (Cinzia Daraio) is a member of the Board of *Scientometrics*.

References

Ahmed, B. S., Amroush, F., & Maati, M. B. (2018). Improving data quality in intelligent eCRM applications. In Mehdi Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (4th ed., pp. 1616–1626). IGI Global.

Angelini, M., Daraio, C., & Urban, L. (2021). A visual analytics environment for the assessment of information quality of performance models. In W. Glänzel, S. Heeffer, P.-S. Chi, & R. Rousseau (Eds.), *The*

18th International Conference on Scientometrics & Informetrics (ISSI2021), 12–15 July 2021 Proceedings (pp. 53–58), ISBN 9789080328228, July 2021.

Angelini, M., Daraio, C., Lenzerini, M., Leotta, F., & Santucci, G. (2020). Performance model's development: A novel approach encompassing ontology-based data access and visual analytics. *Scientometrics, 125*(2), 865–892.

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer International Publishing.

Bors, C., Gschwandtner, T., Kriglstein, S., Miksch, S., & Pohl, M. (2018). Visual interactive creation, customization, and analysis of data quality metrics. *Journal of Data and Information Quality (JDIQ), 10*(1), 1–26.

Cashman, D., Xu, S., Das, S., Heimerl, F., Liu, C., Humayoun, S. R., Gleicher, M., Endert, A., & Chang, R. (2021). CAVA: A visual analytics system for exploratory columnar data augmentation using knowledge graphs. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1731–1741. https://doi.org/10.1109/TVCG.2020.3030443

Cook, K. A., & Thomas, J. J. (2005). Illuminating the path: The research and development agenda for visual analytics (No. PNNL-SA-45230).In *Pacific Northwest National Lab*. (PNNL), WA.

Daraio, C., Bruni, R., Catalano, G., Daraio, A., Matteucci, G., Scannapieco, M., Wagner-Schuster, D., & Lepori, B. (2020). A tailor-made data quality approach for higher educational data. *Journal of Data and Information Science, 5*(3), 129–160.

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics, 108*(1), 441–455.

do Amor Divino Lima, R. S., Davi Oliveira de Araújo, T., Resque dos Santos, C. G., & Serique Meiguins, B., A Visual-Interactive Idiom to Diagnose Missing Data Mechanisms,In *2020 24th International Conference Information Visualisation* (IV), 2020, (pp. 109–113), doi: https://doi.org/10.1109/IV51561.2020.00027.

Ehrlinger, L., Rusz, E., & Wöß, W. (2019). *A survey of data quality measurement and monitoring tools*. https:// arXiv.org/1907.08138

Gschwandtner, T., Aigner, W., Miksch, S., Gärtner, J., Kriglstein, S., Pohl, M., & Suchy, N. (2014). TimeCleanser: A visual analytics approach for data cleansing of time-oriented data. In *Proceedings of the 14th International Conference on Knowledge Technologies AND Data-Driven Business* (pp. 1–8).

He, W., Lei, J., Chu, X., Xie, S., Zhong, C., & Li, Z. (2021). A visual analysis approach to understand and explore quality problems of AIS data. *Journal of Marine Science and Engineering, 9*(2), 198. https://doi.org/10.3390/jmse9020198

Hussain, Md., & Mahmud, I. (2019). PyMannKendall: A python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software, 4*(39), 1556.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 547–554).

Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal, 42*(4), 303–312.

Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., Yu-Shuen, W., & Hong, S. (2018). Steering data quality with visual analytics: The complexity challenge. *Visual Informatics, 2*(4), 191–197.

OpenRefine. (2022). *A free, open source, powerful tool for working with messy data*. Retrived January 1, 2021, from https://openrefine.org/

Song, H., Fu, Y, Saket, B, & Stasko J. Understanding the Effects of Visualizing Missing Values on Visual Data Exploration. In *Proceedings of the 2021 IEEE Visualization Conference* (VIS); Oct. 24 - 29 2021; New Orleans, LA (pp. 161–165), ISBN: 978–1–6654–3335–8

Soylu, A., Giese, M., Jimenez-Ruiz, E., Kharlamov, E., Zheleznyakov, D., & Horrocks, I. (2017). Ontology-based end-user visual query formulation: Why, what, who, how, and which? *Universal Access in the Information Society, 16*(2), 435–467.

Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., Skjæveland, M. G., Hovland, D., Schlatte, R., Brandt, S., Lie, H., & Horrocks, I. (2018). OptiqueVQS: A visual query system over ontologies for industry. *Semantic Web, 9*(5), 627–660.

Sulo, R., Eick, S., & Grossman, R. (2005). DaVis: A tool for visualizing data quality. *Posters Compendium of InfoVis, 2005*, 45–46.

Tukey, J. W. (1977). *Exploratory data analysis*. Springer.

Vielberth, M., Englbrecht, L., & Pernul, G. (2021). Improving data quality for human-as-a-security-sensor. A process driven quality improvement approach for user-provided incident information. *Information and Computer Security, 29*(2), 332–349.